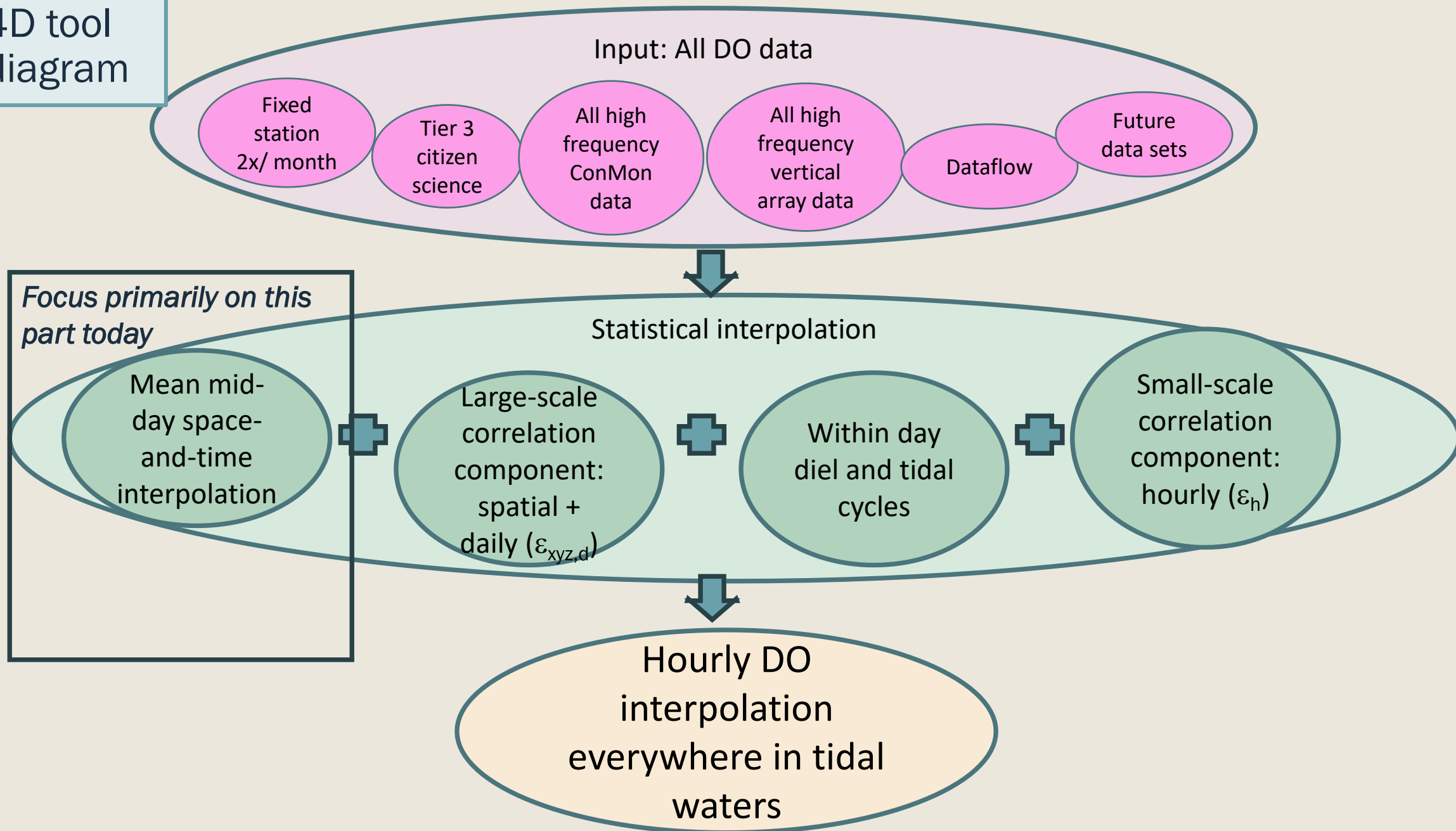


Class Imbalance with 4-D Data Sets and Proposed Solutions

BORG meeting
Feb. 23, 2026

Rebecca Murphy (UMCES at CBP) for Development Team:
Elgin Perry (statistics consultant), Jon Harcum (Tetra Tech),
Breck Sullivan (USGS), and Peter Tango (USGS)

4D tool diagram



Introduction

- A 4-D interpolator strength and challenge is using data sets with very different spatial and temporal scales together.
 - *High frequency temporal data tends to be in shallow waters and can far exceed the count of boat-collected (grab samples) in a year: >99% to <1%.*
 - *The high frequency data is critical to many parts of the interpolation. However, without careful consideration, shallow water DO patterns could be over-represented in the results.*
- The GAM approach was designed to deal with this, to some degree, by fitting relationships between location and DO. But it appears there is a limit to overcoming this data imbalance when data is input at a finer time step than daily.

Explorations have included:

Hourly timestep

In Nov., we presented the idea of using high frequency data at the hourly timestep. Analysis showed this does not lose any information on the shape of the DO distributions.



Insights from the literature

This is a well-known problem in many fields of data analysis called “**class imbalance**.” We will not be outliers if we implement approaches to handle it. Elgin discussed.



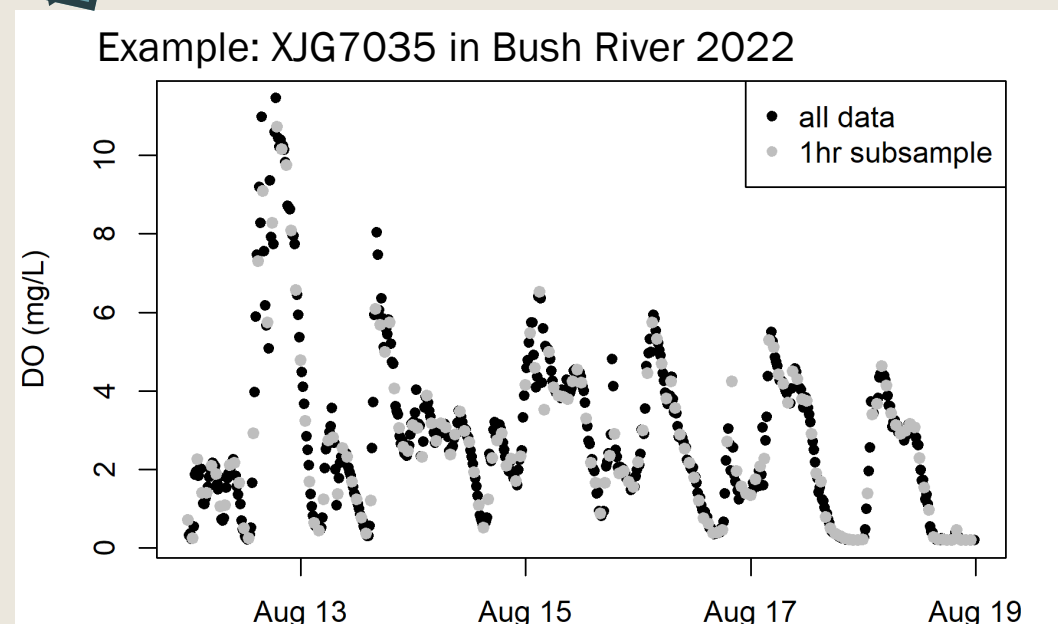
Observation Units/Weighting

All hourly data will still be used in the GAM. However, different weights are applied to the frequency classes before fitting to balance the impact of high frequency-to-grab sample data.

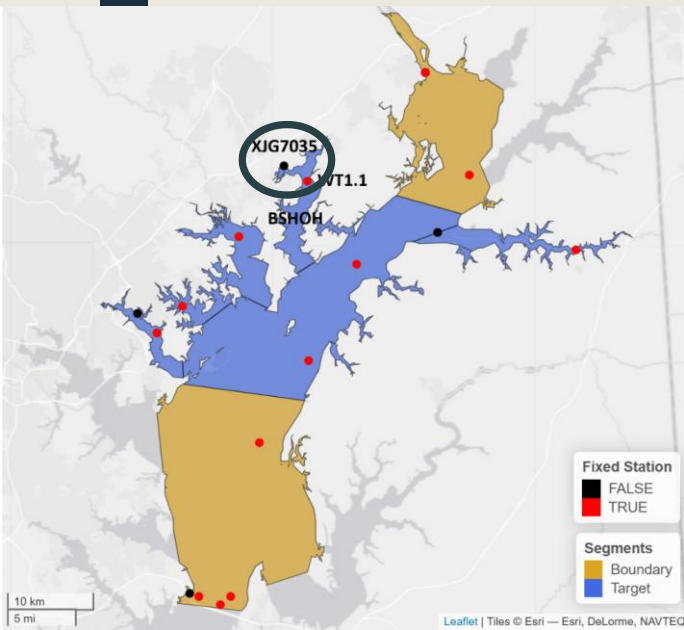


Hourly timestep

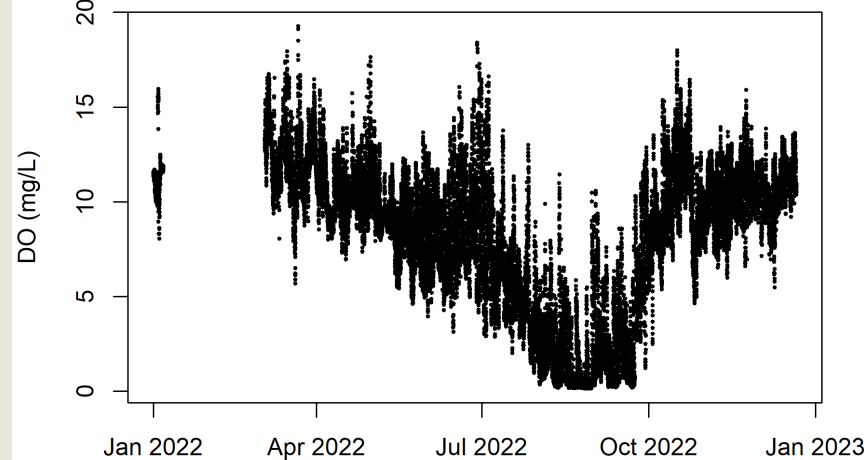
- Recent CAP & BORG discussion (and previous documentation) suggest that an hourly timestep would be appropriate for instantaneous minimum computations.
- Our high frequency data sources provide observations at 15 or 10 minutes.
- **Proposal → Subsample the high frequency data to hourly.** Some reasons:
 - *Input data at a higher frequency than the results would provide little benefit to the interpolation (simply pseudo-replicates).*
 - *Hourly input data helps reduce the class imbalance.*
 - *Why subsample instead of average? Subsampling is preferred to averaging to retain the variance of the input data.*
- Finally, analyses show no difference in criteria-related conclusions between 15 min and 1 hour in our data (next):



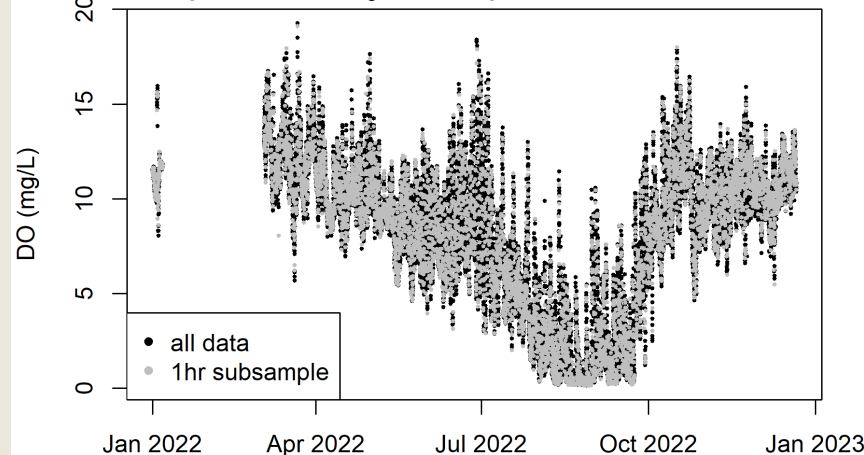
What if ConMon is sub-sampled to hourly?: Example 1



Full data: XJG7035 in Bush River 2022

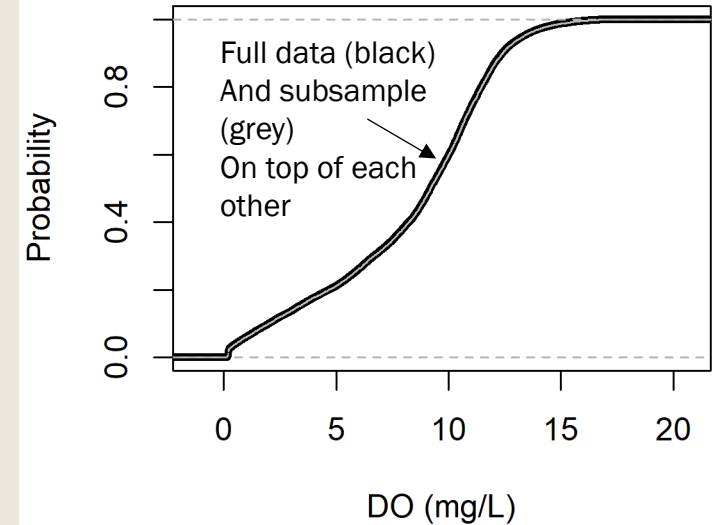


Sub-sampled hourly on top of full data



EDF = Empirical density function

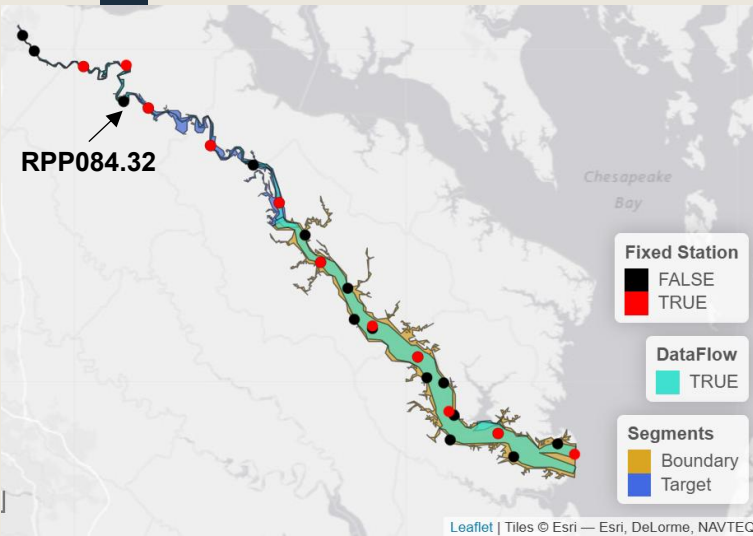
EDF for XJG7035 2022 data



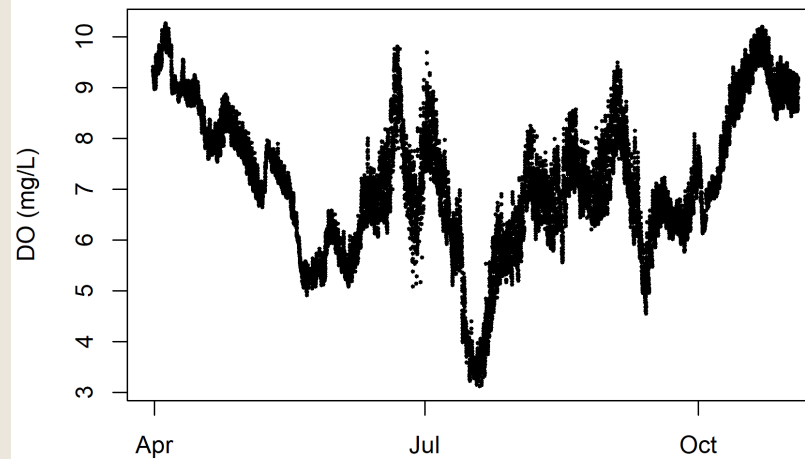
Data set	count	10 th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	28,127	1.97	0.147	0.212
Sub-sample	7,071	1.98	0.146	0.211

These summaries indicate we are not changing the important features of this dataset by sub-sampling this 15 min data to 1 hour.

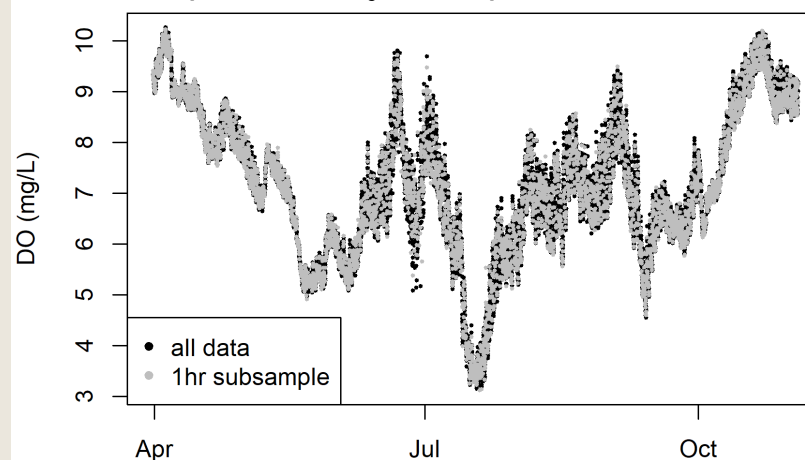
Another example



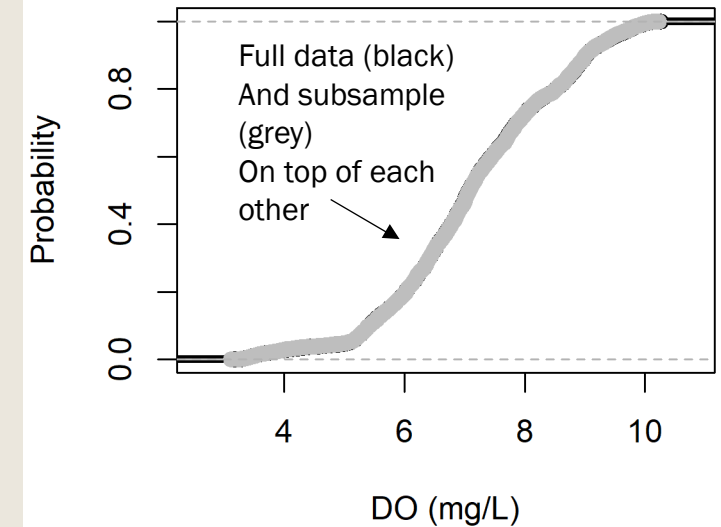
Full data: RPP084.32 in Rappahannock 2022



Sub-sampled hourly on top of full data



EDF for RPP084.32 2022 data



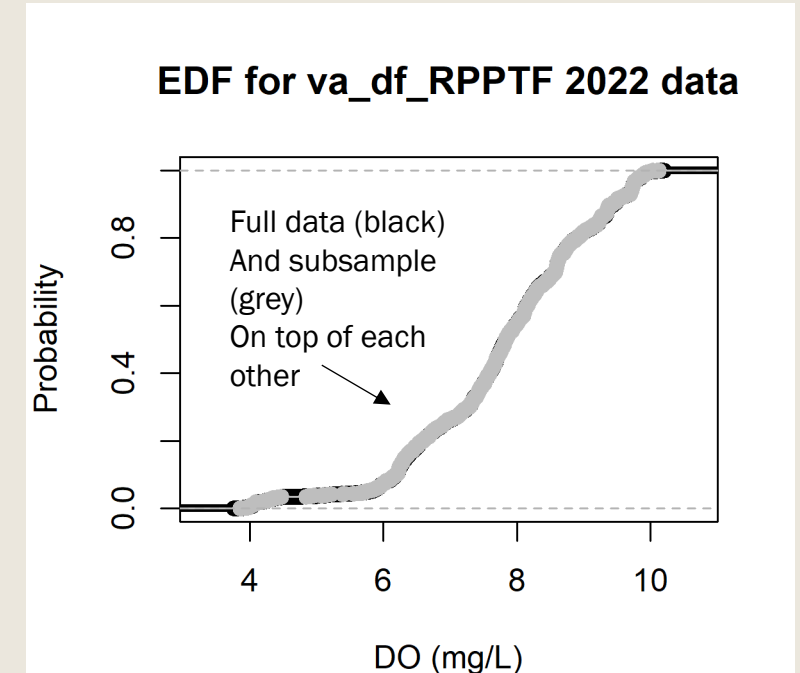
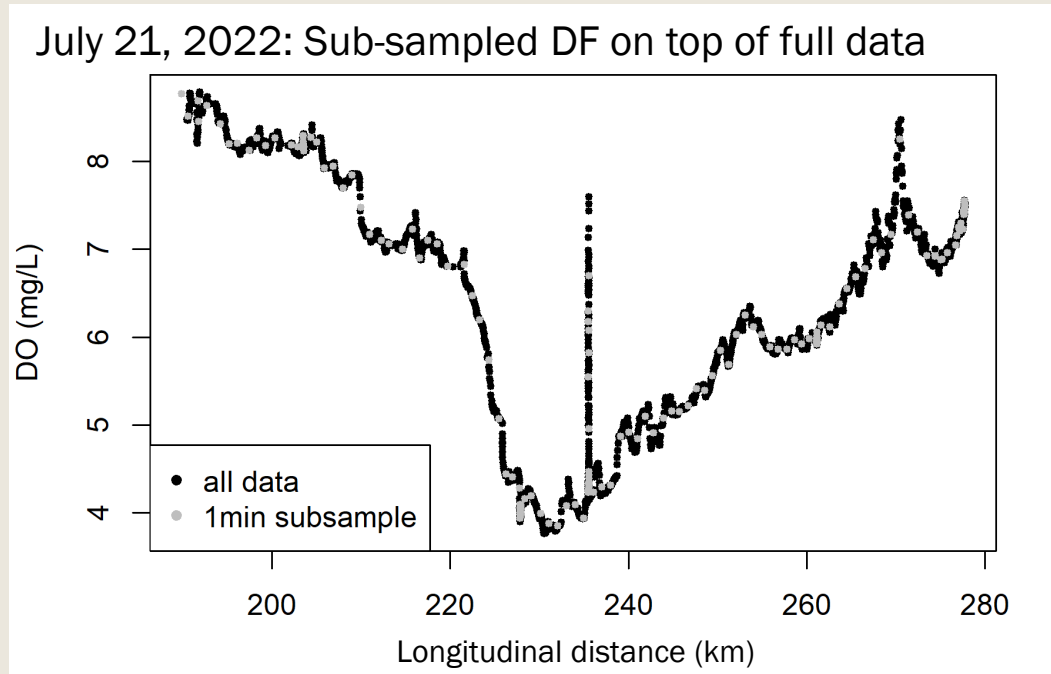
Data set	count	10 th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	20,927	5.42	0.00029	0.0467
Sub-sample	5,316	5.42	0.00038	0.0478

These summaries indicate we are not changing the important features of this dataset by sub-sampling this 15 min data to 1 hour.

Dataflow subsampling

- Dataflow is different from ConMon in that the location is changing as high frequency data is collected.
- To subsample the Dataflow, we are suggesting to retain a sample every 500m. This is to balance having a point or two in almost every grid cell (ranging 100m-1km size).

Subsampling impact: RPPTF dataflow in 2022



Data set	count	10 th percentile	Fraction <3.2 mg/L	Fraction < 5mg/L
All	29,295	6.20	0	0.0374
Sub-sample	1,423	6.19	0	0.0379

Similar conclusion for sub-sampling dataflow.

Take-away: It seems reasonable to include high frequency data at the hourly level (and dataflow at 500m) for the mid-day interpolation.



Hourly timestep

In Nov., we presented the idea of using high frequency data at the hourly timestep. Analysis showed this does not lose any information on the shape of the DO distributions.



Insights from the literature

This is a well-known problem in many fields of data analysis called “**class imbalance**.” We will not be outliers if we implement approaches to handle it. Elgin discussed.



Observation Units/Weighting

All hourly data will still be used in the GAM. However, different weights are applied to the frequency classes before fitting to balance the impact of high frequency-to-grab sample data.



Next focus



Hourly timestep

In Nov., we presented the idea of using high frequency data at the hourly timestep. Analysis showed this does not lose any information on the shape of the DO distributions.



Insights from the literature

This is a well-known problem in many fields of data analysis called “**class imbalance.**” We will not be outliers if we implement approaches to handle it. Elgin discussed.



Observation Units/Weighting

All hourly data will still be used in the GAM. However, different weights are applied to the frequency classes before fitting to balance the impact of high frequency-to-grab sample data.

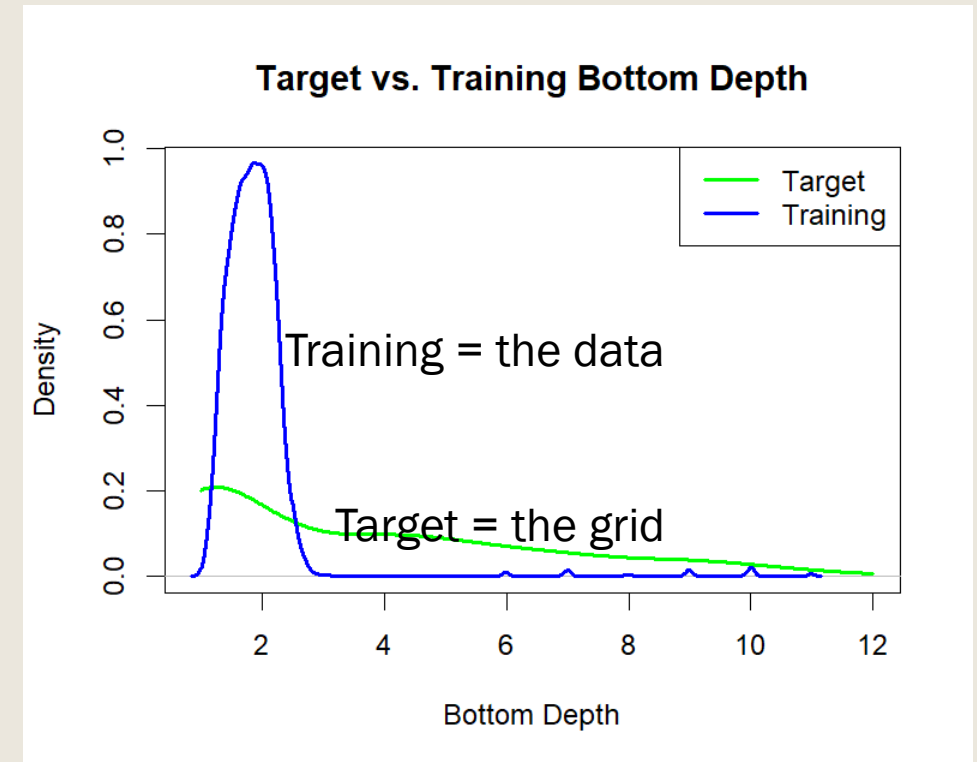


Recall from Elgin's Jan. presentation

Class Imbalance = imbalance between data used for training and data used for prediction.

Identified by comparing the distribution of the independent variables in the training data to the distribution of independent variables in the prediction space or target data.

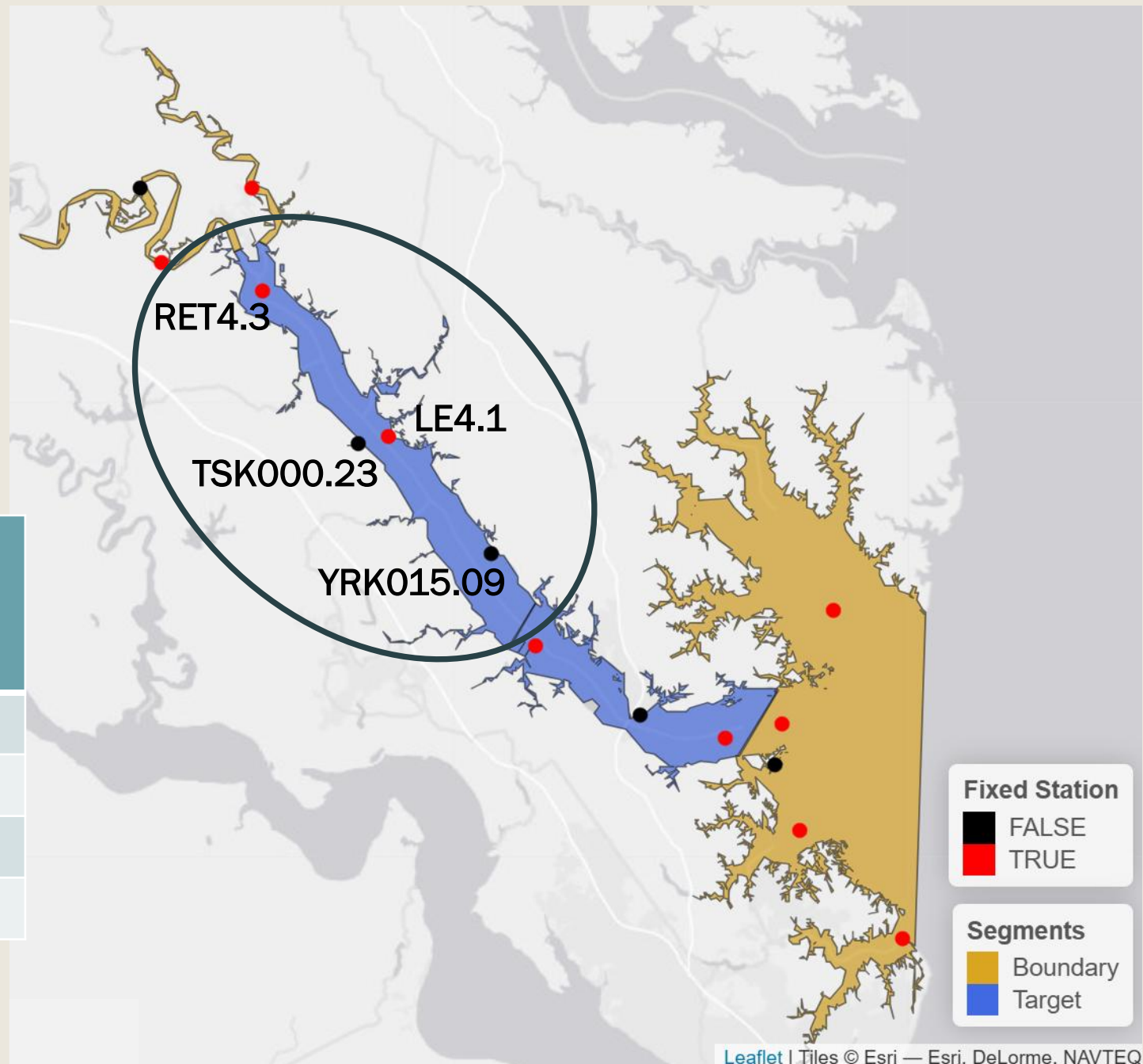
If a region of the training data has a much higher or much lower density of observations than the target data, then **Class Imbalance** exists.



Example

2022 Stations in YRKMH

Station	Median bottom depth (m)*	Count of samples (hour-depts)*	Count of hours with samples*
RET4.3	7	56 (0.33%)	10 (0.06%)
LE4.1	10	87 (0.52%)	10 (0.06%)
TSK000.23	1.8	8,512 (51%)	8,512 (51%)
YRK015.09	1.9	8,146 (48%)	8,146 (49%)

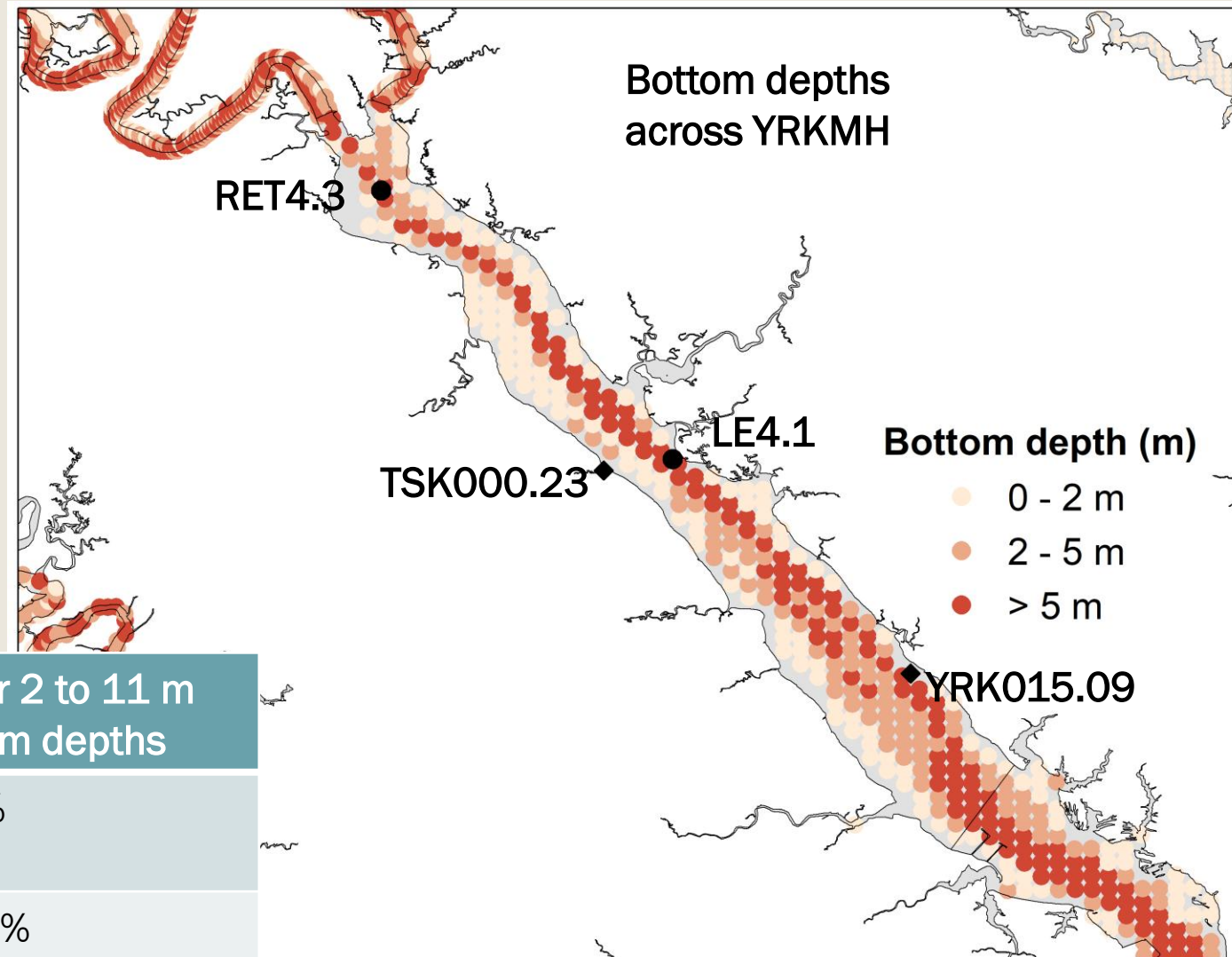


Spatial frequencies in sampling

→ Shallow waters are over-represented in this data.

	Water 0 to 2m bottom depths	Water 2 to 11 m bottom depths
Percent of hourly samples*	99.1 %	0.9 %
Percent of volume	16.8 %	83.2 %

*This includes all depths.



Recall from Elgin's Jan. presentation

For the 4-D Interpolator

Training data =

Fixed Station Data,
Citizen Monitoring
High Frequency Data (ConMon, DataFlow, Vertical Array)

Prediction Space or Target Space =

The interpolation grid

Waterbody Length (100m - 1 km)
Waterbody Breadth (100m - 1 km)
Sample Depth (1 m)
Time (primary 1 day, secondary 1 hour)

Majority Class =

Region with Training data > Target Data

Minority Class =

Region with Training data < Target Data defines

When Class Imbalance becomes a problem.

- Least Squares training adjusts model coefficients to Minimize Prediction Error over all training data.

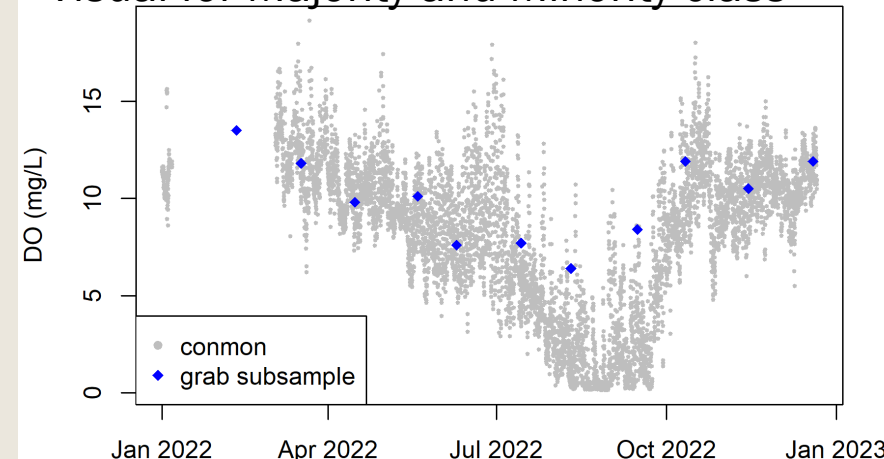
Small improvement x
majority class

>

Large degradation x
minority class

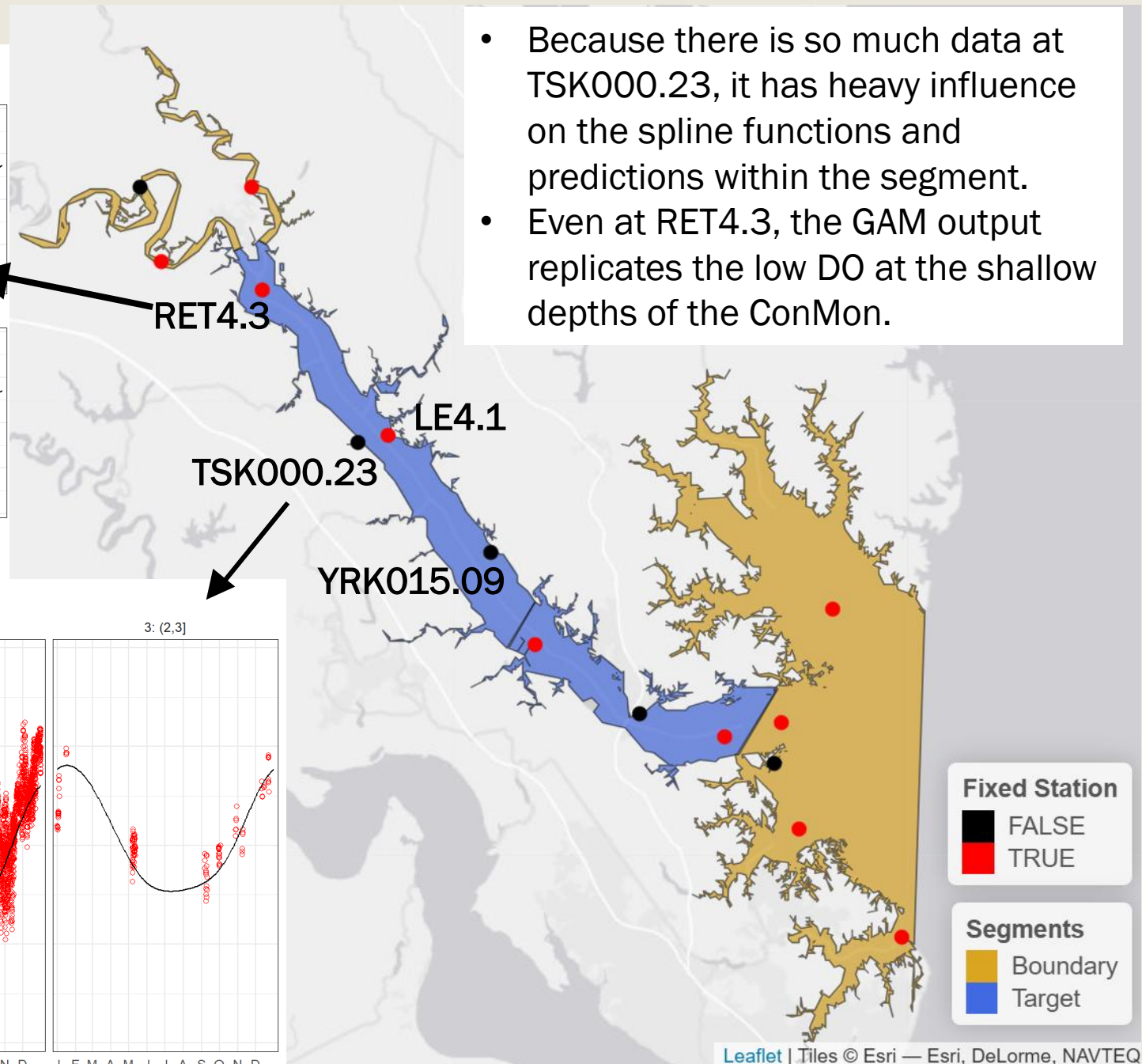
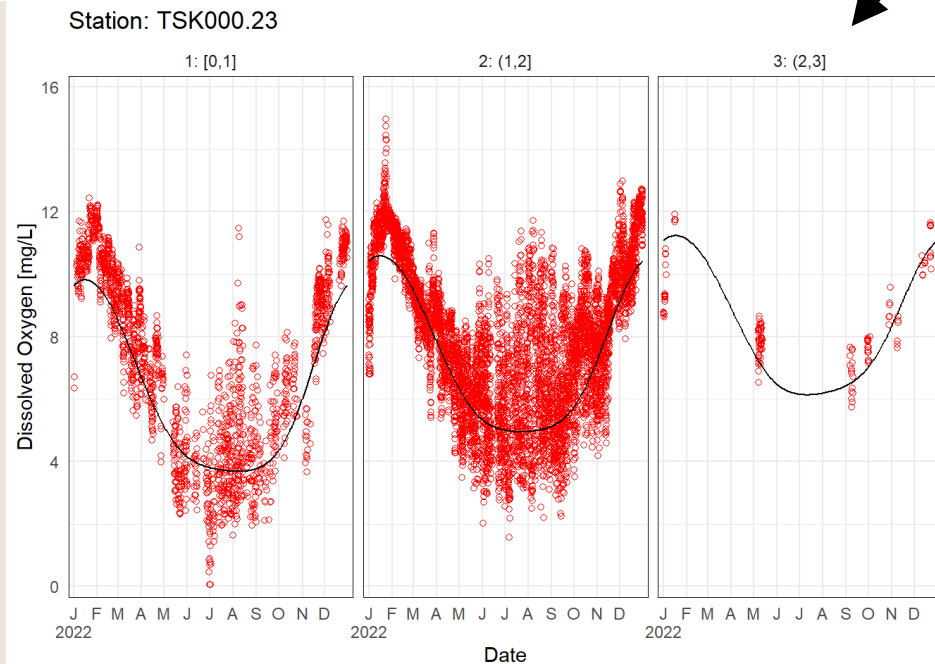
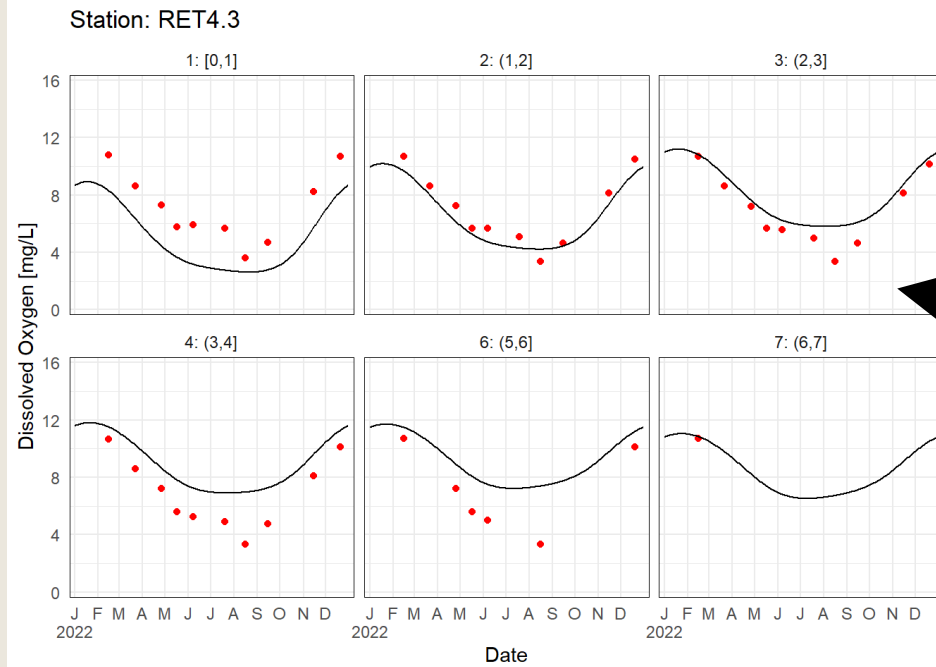
- However, the Minority Class represents a large region in the prediction space. This can lead to poor prediction in a large part of the prediction space.

Visual for Majority and Minority class



Results: no weighting

- Because there is so much data at TSK000.23, it has heavy influence on the spline functions and predictions within the segment.
- Even at RET4.3, the GAM output replicates the low DO at the shallow depths of the ConMon.

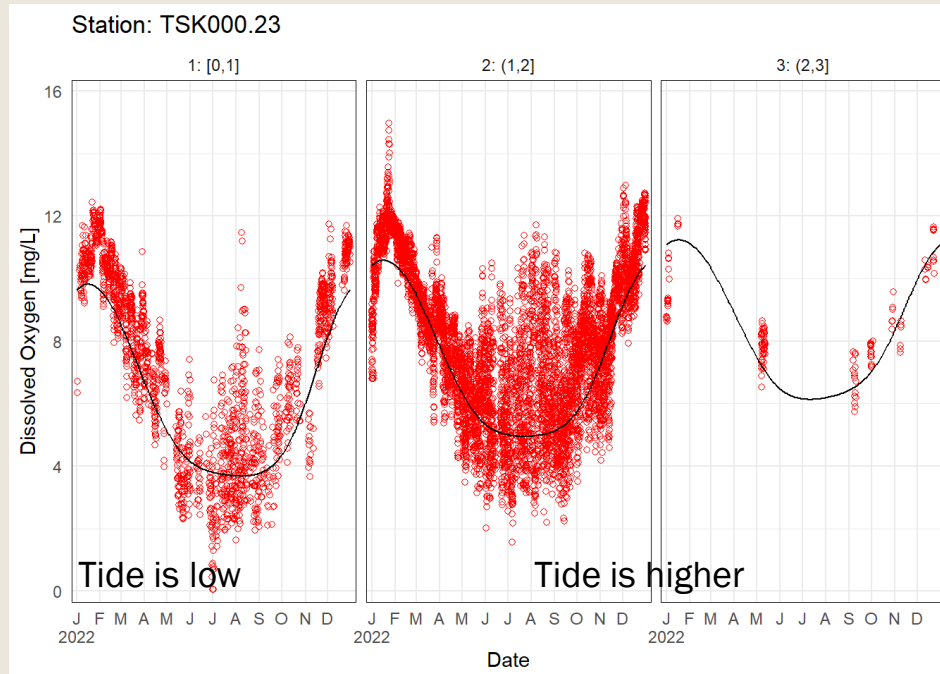


More info

- TSK000.23 is in Taskinas Creek whose depth varies from $\sim 1 - 3$ m.
- This DO pattern might be an important feature for similar creeks, but it has an out-sized impact on the spline fits throughout the York MH due to the large quantity of data.



From Google Maps



We know this situation does not happen in every segment. But looking bay-wide, there are multiple cases of high frequency stations influencing results simply due to the data count imbalance. Any single situation could be explained, but we need the interpolator to work with all data without unique solutions in each segment.

Proposed solution

Think in terms of Observation Units

1) Primary Units are defined by interpolation grid and sampling program –

Station (usually one per grid cell)

Sample depth (1m) or layer

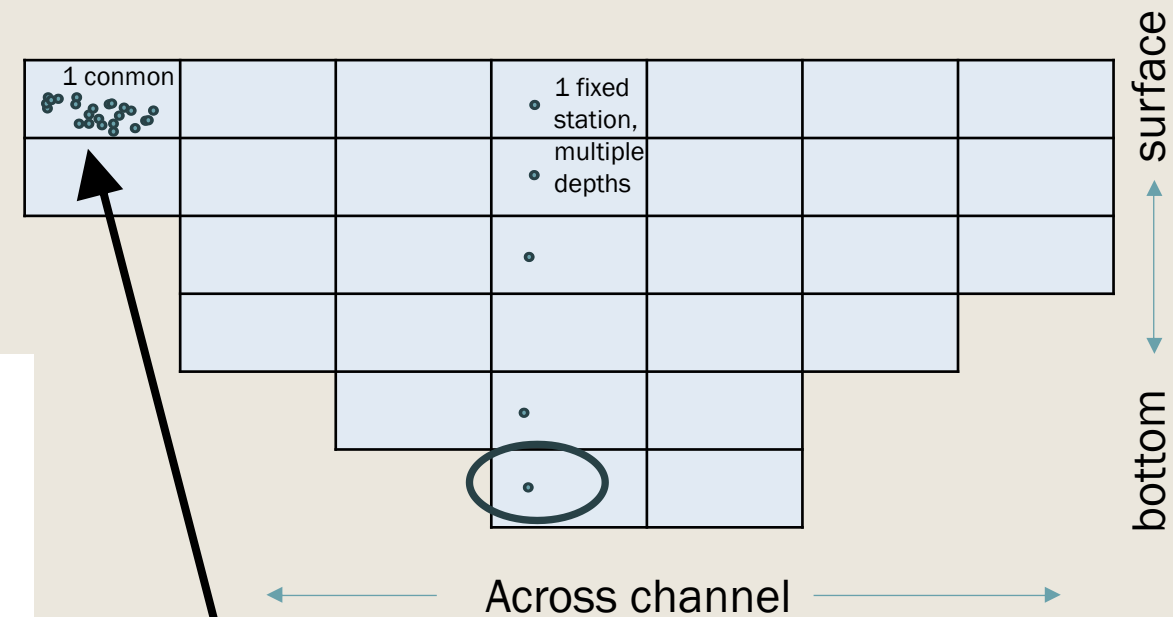
Time (1 day)

→ *For Mid-Day DO + correlated errors*

2) Secondary Units nested within Primary Units –

Time (1 hour)

→ *For within-day and tidal cycles + correlated errors*

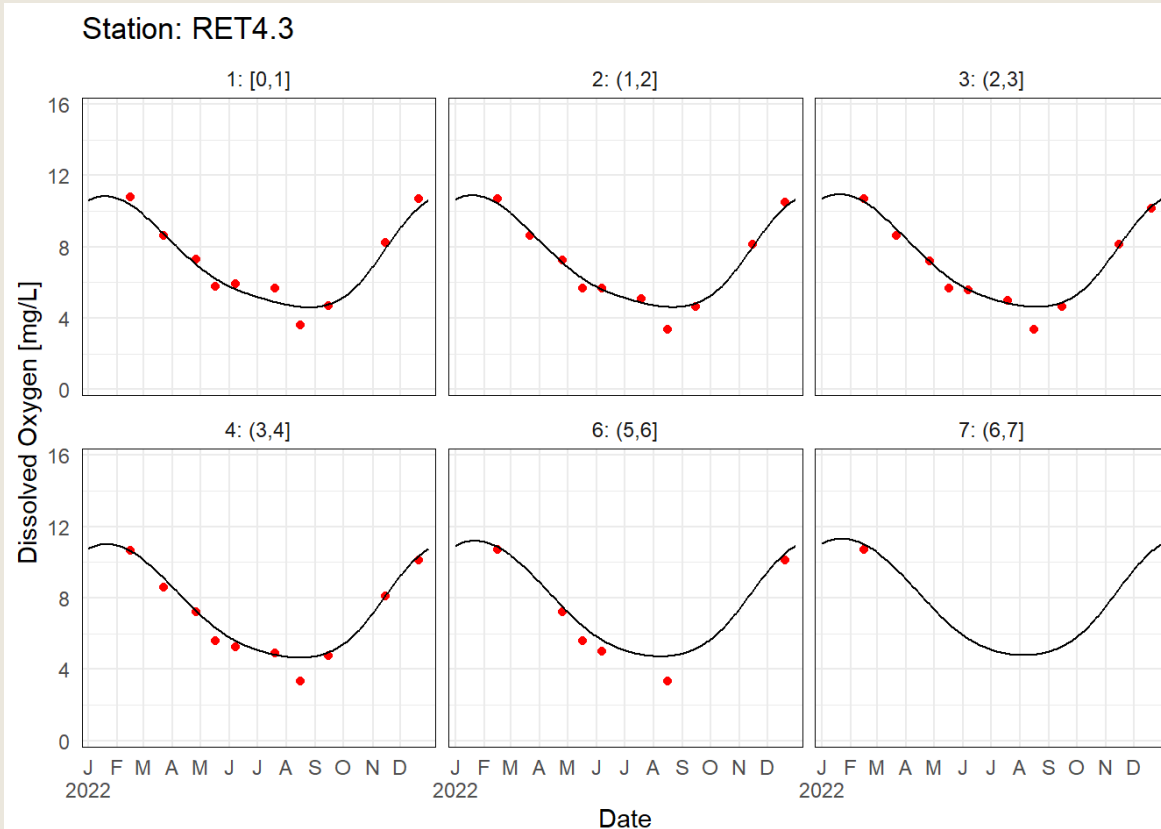


If there is >1 data point in an observation unit, apply a weight to each of the data points so their sum = 1. *This is only for fitting the smooth spline functions.*

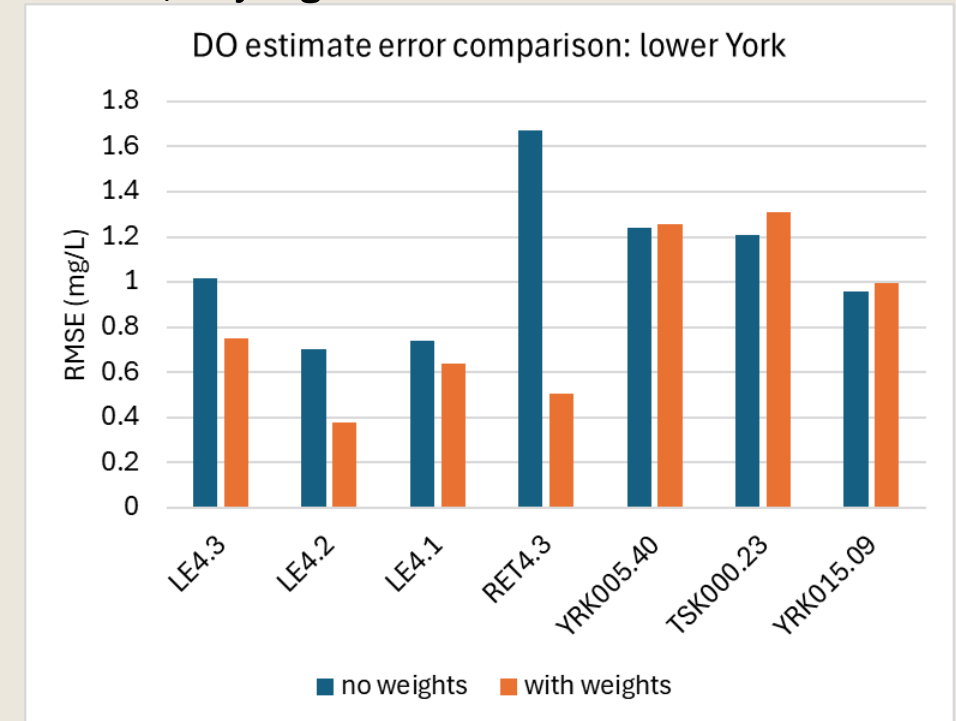
If there is only 1 data point in an observation unit, give it a full weight of 1.

- **Approach: Weight by “Observation Units”.**
 - Each unique “station-date-depth | layer” is one observation unit.
 - If there is >1 sample in a day, each sub-daily sample is given a smaller weight so that the weights on all samples in that day add up to 1.
 - This is **ONLY** for the spline fitting and does not delete any hourly data.
- Keep in mind, for a location with the same conditions as TSK000.23 (shallow water, etc), the interpolation will heavily use that data, even with this approach.

Result at RET4.3 – Daily GAM with weighting fits through the data better



Results in the region: RMSEs decrease at fixed stations, very slight increases at ConMons



Conclusions

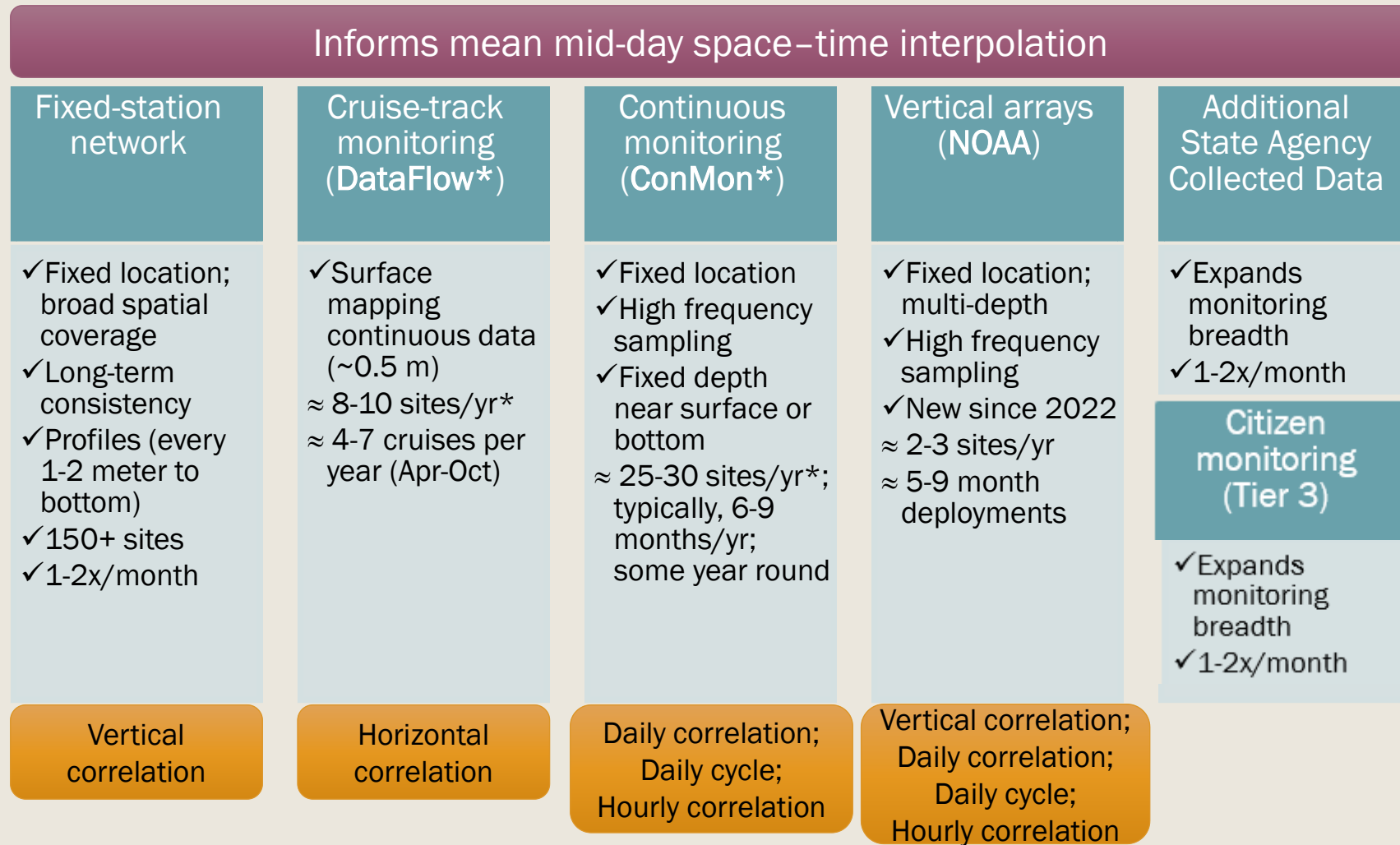
Combining data set types needs to be done thoughtfully to ensure all conditions are interpolated accurately.

- **Hourly subsampling** is sufficient for retaining the nature of the high frequency data, and matches the output likely needed for criteria assessment.
- **Weighting observations** when >1 sample is collected in an “Observation Unit” for the mean mid-day GAM* improves interpolation results across regions. Hourly data will be retained, but are weighted proportionally in fitting the GAM smooth functions.
 - *Comparison was done in every segment, both for 2022 and 2016.*
- This challenge is not unique to the 4-D interpolator. Likely we’d be looking at this challenge in some way if just aggregating the raw data for violations.

*Other parts of the 4-D interpolator uses **station-hour-depth** as the “Observation Unit.”

extras

Data Landscape

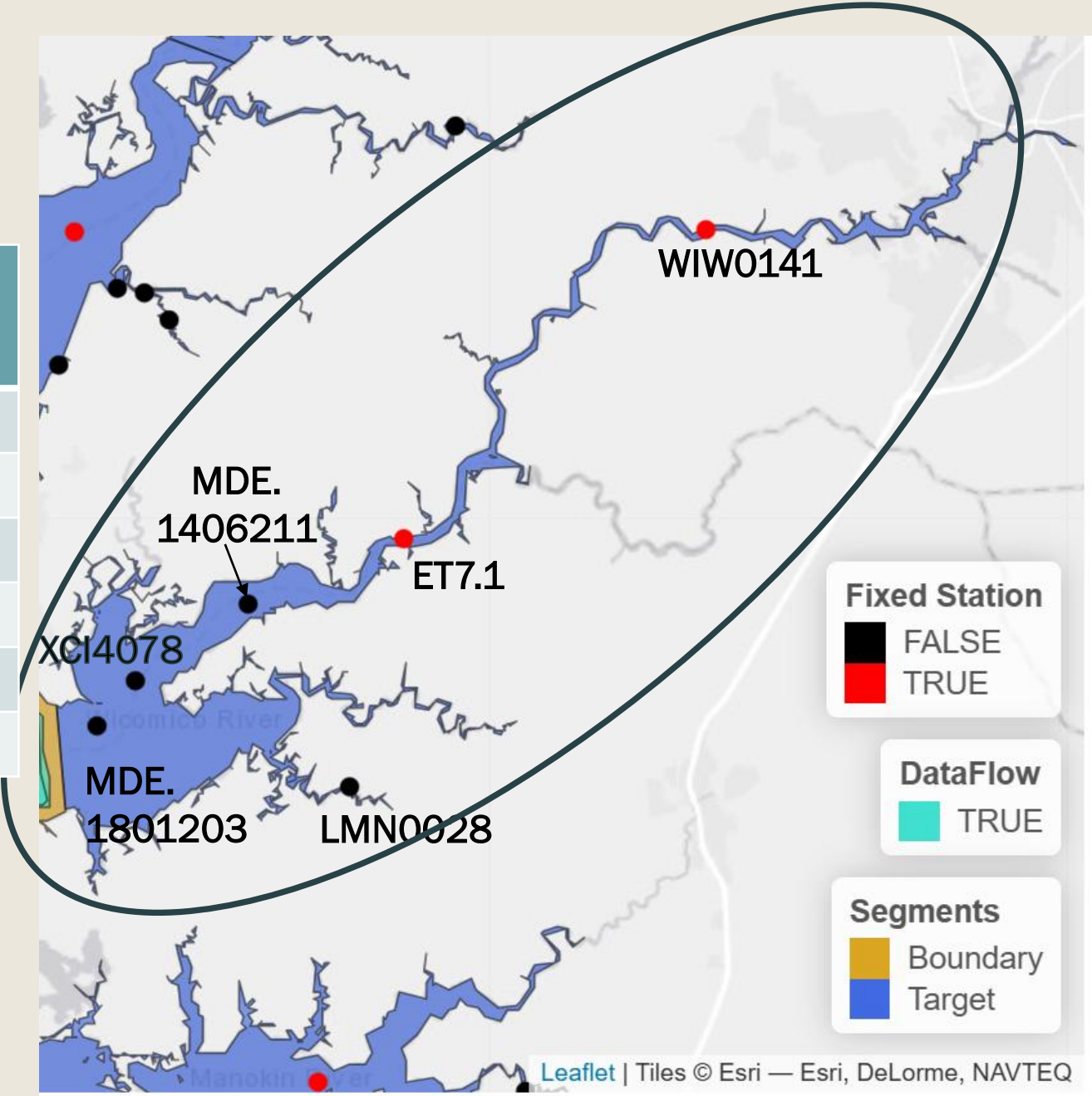


*Sites commonly rotated about every 3 years to broaden coverage

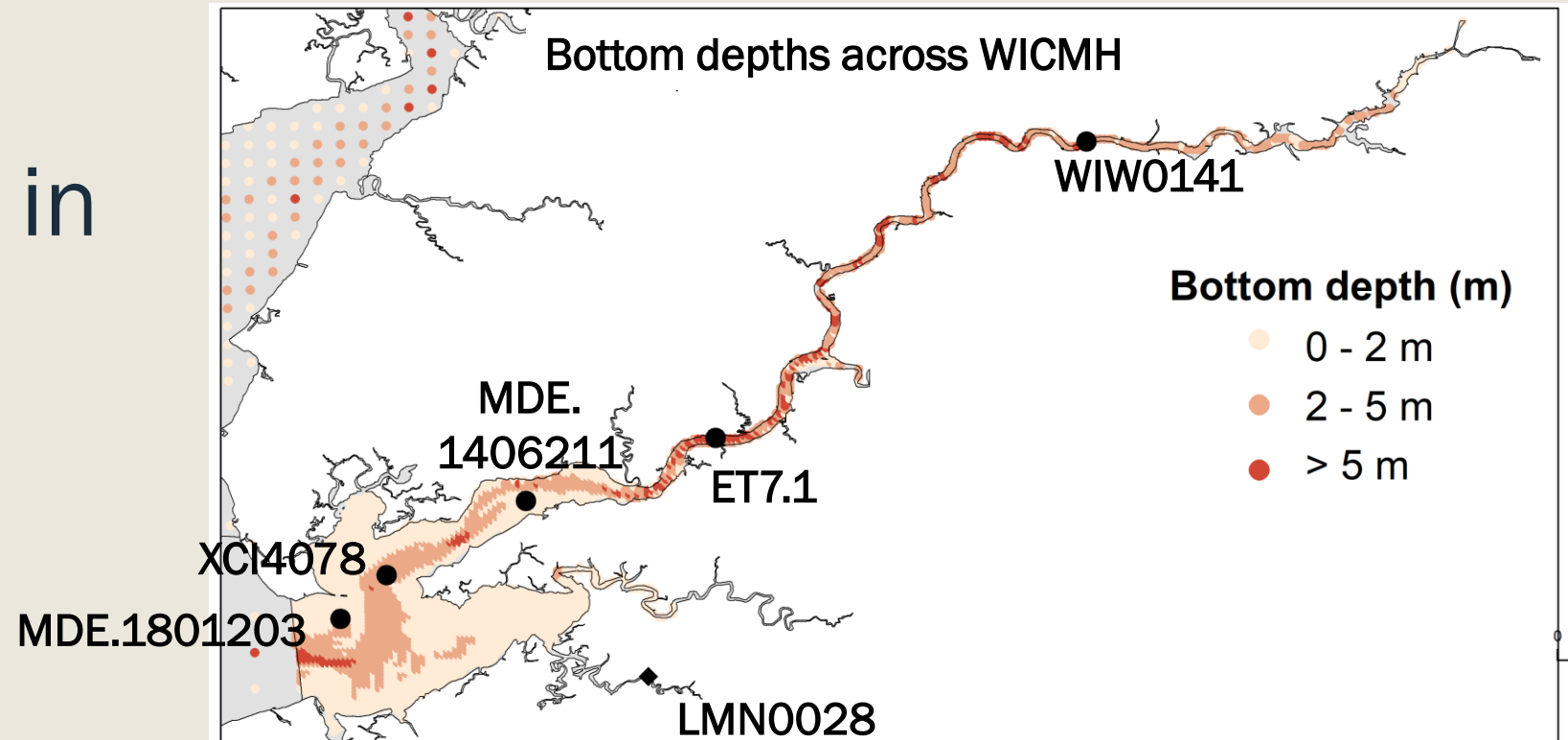
Example 2

2022 Stations in WICMH

Station	Median bottom depth (m)*	Count of samples (hour-depts)*		Count of hours with samples*	
WIW0141	5.7	10	(0.13%)	10	(0.13%)
ET7.1	7.4	79	(1.0%)	12	(0.15%)
MDE.1406211	1.7	7	(0.088%)	6	(0.077%)
XCI4078	3.7	45	(0.57%)	12	(0.15%)
MDE.1801203	2.2	9	(0.11%)	5	(0.064%)
LMN0028	1.3	7,763 (98%)		7,740 (99%)	



Spatial frequencies in sampling

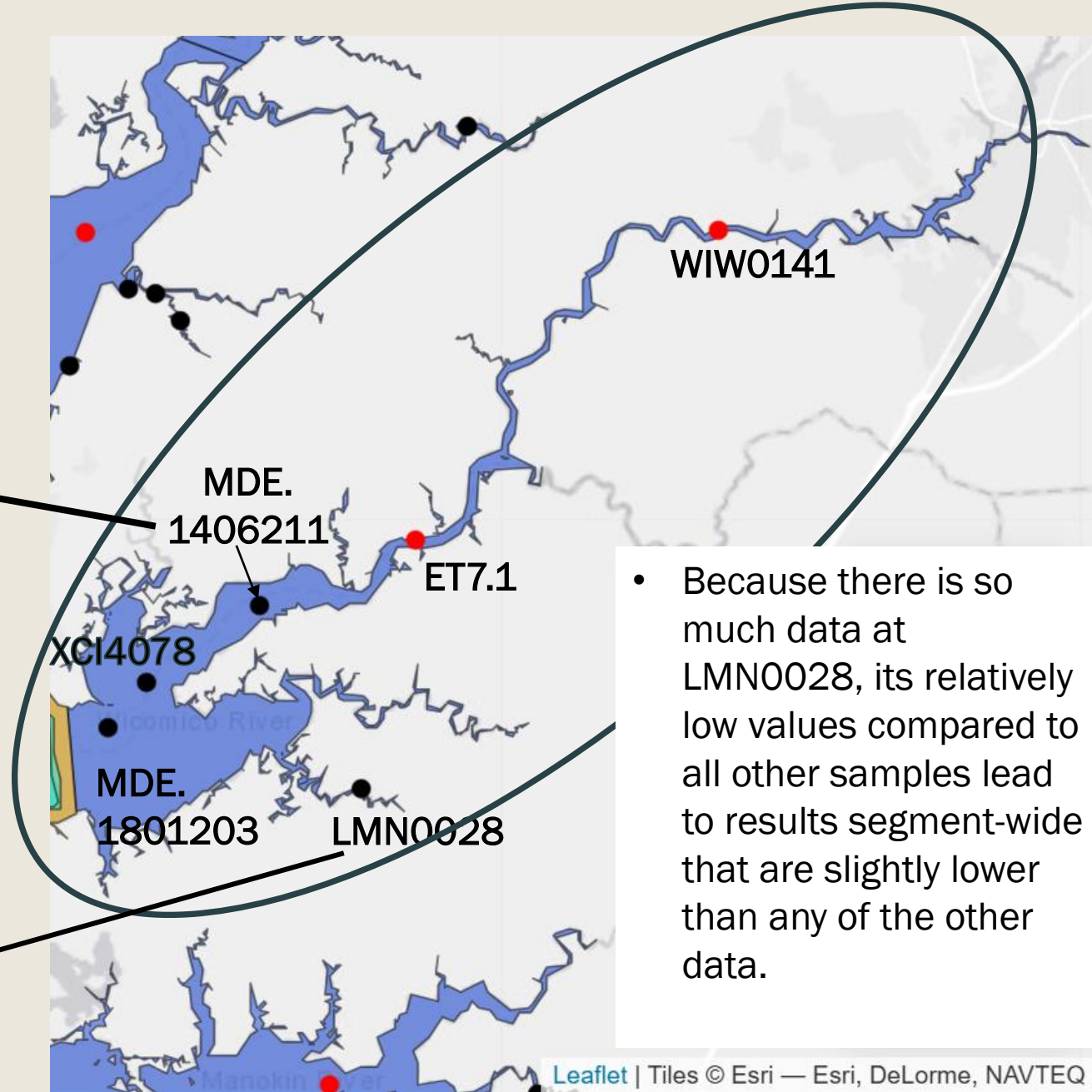
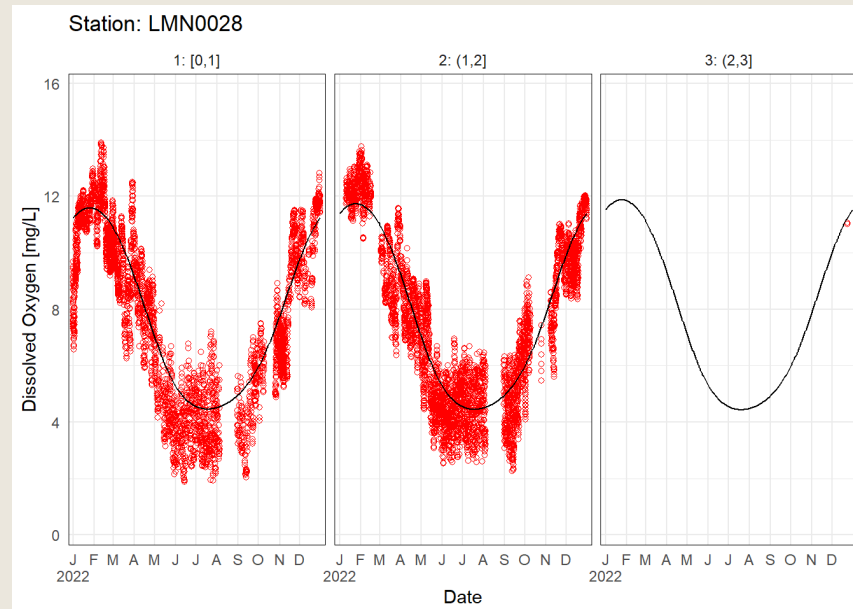
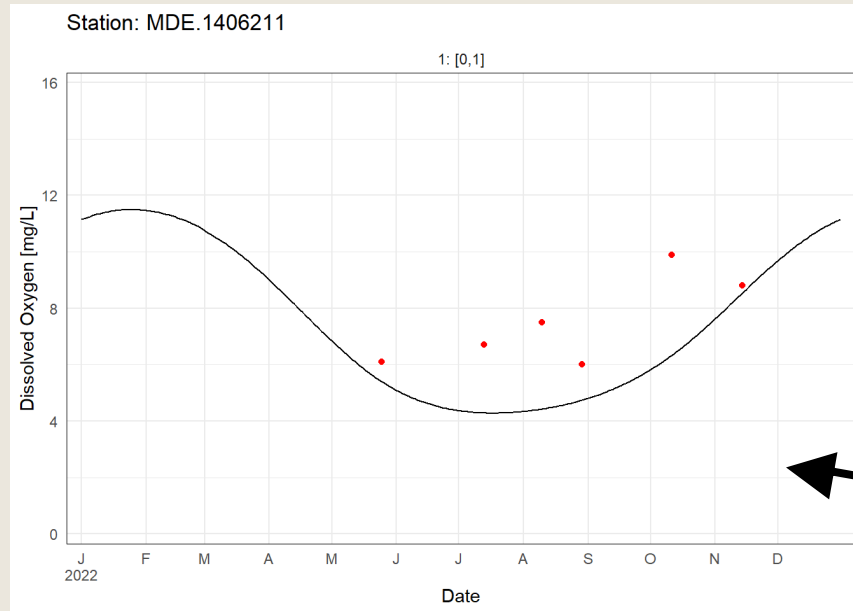


	Water 0 to 2m bottom depths	Water 2 to 12 m bottom depths
Percent of hourly samples*	98.2 %	1.8 %
Percent of volume	51.7 %	48.3 %

→ Shallow waters are over-represented in this data, but not as extremely as YRKMH.

*This includes all depths.

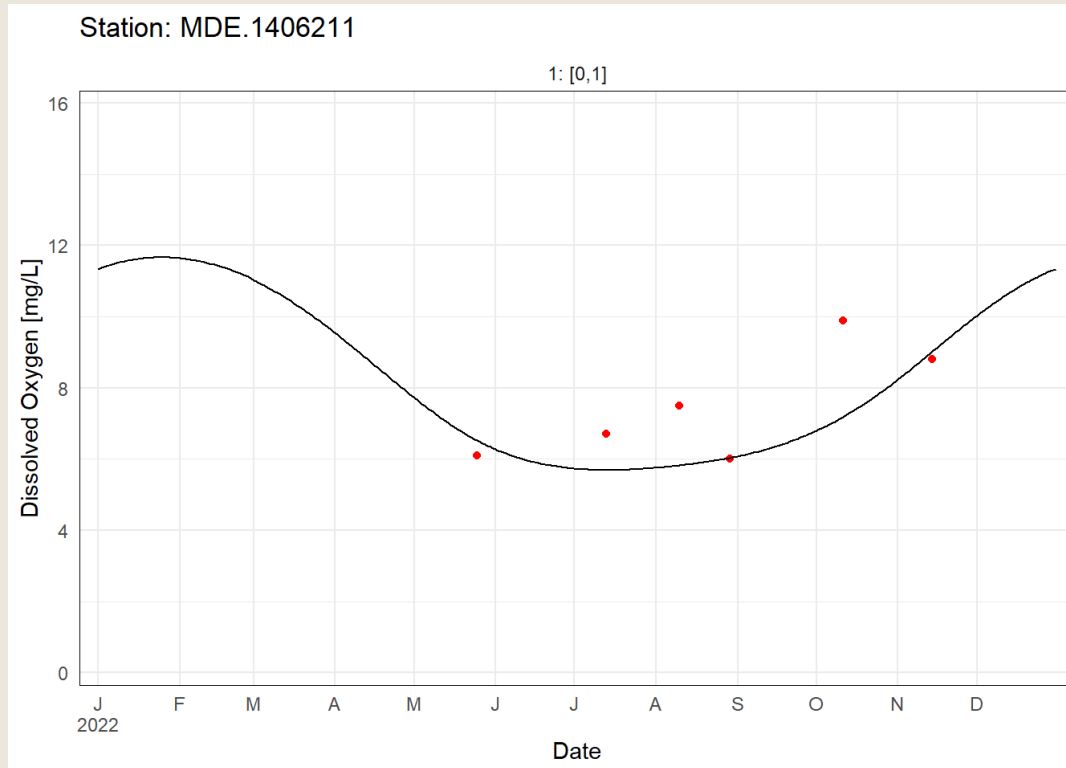
Results: no weighting



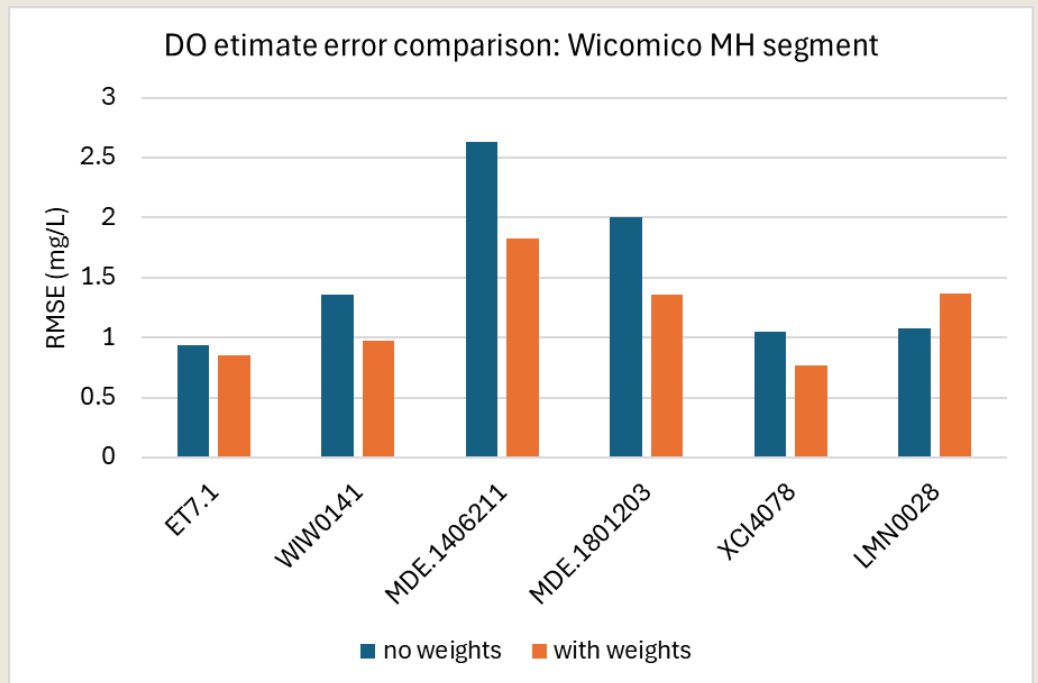
- Because there is so much data at LMN0028, its relatively low values compared to all other samples lead to results segment-wide that are slightly lower than any of the other data.

- Approach: Weight by “Observation Units”.
 - Each unique “station-date-depth | layer” is one observation unit.
 - If there is >1 sample in a day, each sub-daily sample is given a smaller weight so that the weights on all samples in that day add up to 1.
 - This is **ONLY** for the spline fitting and does not delete any hourly data.
- Keep in mind, for a location with the same conditions as LMN0028 (shallow water, etc), the interpolation will heavily use that data, even with this approach.

Result at MDE.1406211 – Daily GAM with weighting fits min better



Results in the region: RMSEs decrease at fixed stations, increases at ConMon



Temporal frequencies in sampling

- It is a huge success that these ConMon stations cover almost every hour of the year!
 - *That teaches us a lot about shallow water DO year-round and we need this information in the 4-D interpolator.*
- However, if the ConMon data are not representative of the deeper waters, we want to be sure they impact the interpolation in only the right places and times.

For YRKMH data in 2022

	One or both ConMon stations	Sampled by either Fixed station
Count of hours with samples	8,724 hours	20 hours
Percent of hours/year	99.6 %	0.228 %

Purpose: Build a tool for more complete criteria assessment

DO criteria that currently can be evaluated with existing approaches and data

Table 1. Chesapeake Bay dissolved oxygen criteria.

Designated Use	Criteria Concentration/Duration	Protection Provided	Temporal Application
Migratory fish spawning and nursery use *	7-day mean ≥ 6 mg liter ⁻¹ (tidal habitats with 0-0.5 ppt salinity)	Survival/growth of larval/juvenile tidal-fresh resident fish; protective of threatened/endangered species.	February 1 - May 31
	Instantaneous minimum ≥ 5 mg liter ⁻¹	Survival and growth of larval/juvenile migratory fish; protective of threatened/endangered species.	
	Open-water fish and shellfish designated use criteria apply		June 1 - January 31
Shallow-water bay grass use	Open-water fish and shellfish designated use criteria apply		Year-round
Open-water fish and shellfish use	30-day mean ≥ 5.5 mg liter ⁻¹ (tidal habitats with 0-0.5 ppt salinity)	Growth of tidal-fresh juvenile and adult fish; protective of threatened/endangered species.	Year-round
	30-day mean ≥ 5 mg liter ⁻¹ (tidal habitats with >0.5 ppt salinity)	Growth of larval, juvenile and adult fish and shellfish; protective of threatened/endangered species.	
	7-day mean ≥ 4 mg liter ⁻¹	Survival of open-water fish larvae.	
	Instantaneous minimum ≥ 3.2 mg liter ⁻¹	Survival of threatened/endangered sturgeon species. ¹	
Deep-water seasonal fish and shellfish use	30-day mean ≥ 3 mg liter ⁻¹	Survival and recruitment of bay anchovy eggs and larvae.	June 1 - September 30
	1-day mean ≥ 2.3 mg liter ⁻¹	Survival of open-water juvenile and adult fish.	
	Instantaneous minimum ≥ 1.7 mg liter ⁻¹	Survival of bay anchovy eggs and larvae.	
	Open-water fish and shellfish designated-use criteria apply		October 1 - May 31
Deep-channel seasonal refuge use	Instantaneous minimum ≥ 1 mg liter ⁻¹	Survival of bottom-dwelling worms and clams.	June 1 - September 30
	Open-water fish and shellfish designated use criteria apply		October 1 - May 31

*Note a 30-day mean 6 mg/L MSN value is evaluated for purpose of the WQ indicator.

¹ At temperatures considered stressful to shortnose sturgeon (>29°C), dissolved oxygen concentrations above an instantaneous minimum of 4.3 mg liter⁻¹ will protect survival of this listed sturgeon species.